Memo

To: Ron Pi, State Bar of California

From: Mary J. Pitoniak, Independent Consultant

Date: August 4, 2017

Re: Evaluation of Standard Setting Conducted for State Bar of California

Overview

This memo represents my evaluation of the standard setting activities for the California Bar Exam conducted by Chad Buckendahl of ACS Ventures for the State Bar of California. Some of the observations contained herein were also provided in my memo of May 8, 2017, regarding the plans for the workshop prior to its being conducted.

This memo is framed in terms of three general sources of validity evidence: procedural, internal, and external (Kane, 1994, 2001; see also Hambleton & Pitoniak, 2006). Also included in the Appendix are comments related to Hambleton's (2001) *Criteria for Evaluating a Performance Standard Setting Study*, many of which overlap with the Kane criteria.

The report referred to in this memo is Buckendahl (2017, July 28). I had reviewed a draft report as well, but have updated this memo to reflect the content of the final version that incorporated suggestions provided by me and other reviewers.

Procedural Evidence

According to Kane (2001), "the fact that a standard setting study has employed an apparently sound procedure in a thorough and systematic way, and has where possible, included various checks on the consistency and reasonableness of the results encourages us to have faith in the results" (p. 68). Procedural evidence from this meeting is grouped into six areas: selection of procedures, explicitness, practicability, implementation of procedures, panelist feedback, and documentation. Panelist feedback is discussed within each of the areas as it appropriate (e.g., did their comments indicate a lack of understanding of a given activity), as well in a separate section.

1) Selection of Procedures

Neither the proposal nor the plan had indicated why the Analytic Judgment method was chosen. The report indicates that the method is appropriate for the characteristics of the exam; a footnote mentions that alternative methods that also involve ratings of candidate work include the Paper Selection and Body of Work methods, but does not describe why they were not chosen.

In a discussion with Dr. Buckendahl before the study I had asked if he had considered other methods. He said that the Extended Angoff method (see Hambleton & Plake, 1995) would have been problematic because of

the inconsistency of the scoring rubrics across the essays and performance tasks (referred to collectively as questions in this memo). In the Extended Angoff method, panelists review the items and the scoring rubric and then provide an estimate of the score that the minimally competent candidate would get on each item. The cutscore is obtained by summing the scores across items. Because of the inconsistent rubrics with the Bar Exam, there would presumably be a risk of panelists getting anchored on one score point and having difficulty starting each item afresh.

As the report noted, another method used for tests comprised entirely of constructed-response items is the Body of Work method (Kingston, Kahl, Sweeney, & Bay, 2001). Panelists review and rate sets of work from the pool of examinees, which in the case of the California Bar Exam would be complete sets of responses to the essays and performance tasks for a group of examinees. Each "body of work" is rated as falling into one of the performance categories. The method involves several rounds; initially panelists rate sets of work from across the score distribution (although those scores are not known to them). After finding the area in the score range where the boundary between the performance categories is seen as likely to fall, additional review activities are undertaken to pinpoint its location. The cut score is then located analytically using logistic regression (Kingston et al, 2001) or through averaging of ratings (Olson, Mead, & Payne, 2002). Given the number of questions administered to each examinee, the Body of Work method may have proved problematic for the Bar Exam.

The choice of the Analytic Judgment method for use with the California Bar Exam seems reasonable. However, as noted later in this memo, I would have made different decisions about the specifics of its implementation in this case.

2) Explicitness

Explicitness refers to the extent to which the standard setting activities were clearly and explicitly described in advance of implementation (van der Linden, 1995). As I noted in my May 8 memo, my concerns about the explicitness of the workshop plans fell into two categories, as described in the following sections.

Lack of One Comprehensive Document

Planning information was spread across multiple documents.

- *Proposal to [sic] a Standard Setting Study for the California Bar Exam*, dated March, 2017
- *California Bar Standard Setting Plan*, dated March 24, 2017
- *California Bar Exam Standard Setting Workshop Agenda—DRAFT*, dated March 24, 2017, and revised version dated April 25, 2017
- *Standard Setting Workshop* PowerPoint presentation for call with deans, dated April 25, 2017

Ideally, the information about how the workshop would be conducted would have been in one comprehensive, detailed document. Such a document would have facilitated my review and evaluation of all of the different design designs that have been made. It would also have enhanced my ability to do a better evaluation of the procedural validity of the workshops now, after they have been conducted. As it stood I needed to check three or four different sources for information.

Lack of Detail

As I noted in an email on March 14, 2017 to Leah Wilson (Chief Operating Officer) after reviewing the proposals, to be able to evaluate evidence for procedural validity, the design, or plan, has to be thorough enough so that a reviewer could evaluate the extent to which the procedures were carried out as intended. Generally plans are detailed enough that they can be used as the basis for the technical report. An additional consideration is whether a future study could use the same procedures using the technical report as a basis, thus ensuring consistency across implementations.

I had thought that perhaps the proposals provided were less detailed since they were a general outline of the work. However, the subsequent documents did not provide additional detail in the following areas:

- Targets for panelist demographics, beyond differentiation between recently licensed attorneys and experienced ones who supervise entry-level attorneys. For example, targets could be set for gender, area of the state, etc., and the characteristics obtained for the panel would be compared against them.
- Description of any materials to be sent in advance, such as the agenda, purpose of the workshop, etc.
- More detailed description of the exam materials (content assessed, item types, etc.) and the rubrics
- Justification for why typically-used parts of the process are not being incorporated, such as feedback to and discussion among panelists, impact data, etc. If part of the reason is a restriction on the length of the workshop, that should be made clear.
- Indication of the type of information that will be included in the technical report (to be sure it is collected during the study)
- More information about the validity criteria that would be used for evaluating the results of the study

3) Practicability

Under the criterion of practicability, the following characteristics of the method are summarized: the difficulty of implementing the methods and the feasibility of data analysis, and the degree to which the methods are credible and interpretable to laypeople (Berk, 1986).

Difficulty of Implementation and Feasibility of Data Analysis

I will discuss the criteria of the difficulty of implementing the method and the feasibility of data analysis together since they are closely related. Berk (1986) notes that the steps required to implement the method need to be clear, understood by the panelists, and able to be executed in a reasonable amount of time. He also states that the method should be easy to compute.

I noted above that the activities to be conducted in the workshop could have been explained in more detail. As a result, the steps needed to implement the method, while rather straightforward, were not made clear in the planning documents. Likewise, the training slides could have been more detailed. Panelists did express some confusion in their responses to open-ended questions about how the ratings would be transformed into cut scores (see the Panelist Feedback section).

In general, the workshop could have benefited from being longer so that more time could be provided for both training and provision of operational ratings. As described in the Panelist feedback section, some panelists indicated that there was not enough or barely enough time for some activities, with their responses to open-ended questions providing more detail.

In terms of data analysis, because other than for the practice round and the first essay question there was only one round of ratings, there was not extensive compilation of ratings during the meetings and thus no data analyses conducted in real-time. Analyses were instead conducted after the meeting and were straightforward. The approaches taken to (a) calculating individual and panel cutscores, (b) relating them to the full score scale, and (c) estimating impact are clearly described in the report.

Credibility and Interpretability

The Analytic Judgment method is relatively simple to explain to panelists. The main task was explained to panelists in two bullets on the orientation slides (slide 17):

- Broadly classify papers into three categories (Not Competent, Competent, Exceeds) using Performance Level Descriptor
- Refine classification decisions to identify the two best Not Competent and the two worst Competent

However, at the workshop I observed that panelists frequently asked questions about how the cut score would be determined, and that was not clearly explicated in the slides—i.e., that the scores assigned to the two papers they identified as the best Not Competent and the two worst Competent would be averaged to get the cut score for each question, after which the cut scores would be summed across the essays and performance task to yield the panelist cutscore, and then panelist cut scores would be averaged to get the recommended panel cut score.

If clear and thorough communication is employed when describing the passing score resulting from the workshop, it is likely that the Analytic Judgment method would appear creditable and interpretable to both policymakers and lay people. The relationship of the scores assigned to the responses and the ratings made by panelists would need to be clearly explicated, followed by a description of how ratings are compiled to obtain panelist-level and panel-level recommended cut scores.

4) Implementation of Procedures

For Kane (1994), an important source of procedural validity evidence is the degree to which the selection of panelists, training of panelists, definition of the performance standard, and data collection are implemented in a systematic and thorough fashion.

Selection of Panelists

In terms of panelists, as outlined above under the explicitness criterion, I noted my concern when reviewing the plan that the only target for panelist demographics was length of practice, specifically the number of recently licensed attorneys vs. experienced ones who supervise entry-level attorneys (the report also noted that some panelists had been selected to represent the Faculty/Educator category—panelists who are employed at a college or university). Generally panelists are selected in relation to other criteria as well, such as gender, area of the state, etc., and the characteristics obtained for the panel are then compared against them (Hambleton & Pitoniak, 2006). As far as I know, no specific targets were set in advance for any of the criteria, though the report does report additional pieces of demographic information. Overall, the total number of panelists was sufficient for this activity. Raymond and Reid (2001) recommend at least 10–15 panelists, and 20 panelists participated in this workshop.

A decision faced during panelist selection for any licensing exam is whether to include panelists who have previously been closely involved with the examination. Raymond and Reid (1991) discuss the advantages and disadvantages of including what they term "members of the existing examination committee" (p. 134). Advantages include sensitivity to the factors that influence item performance, a genuine interest in the process, and possession of knowledge of the items and content specifications: "in short, examination committees are experts who know what to expect of items and examinees" (p. 134). However, they note that some agencies have policies that specifically exclude such individuals from the standard setting process so that the standard-setting group is totally independent of the test development process. The examination committee members may also be viewed as having expectations or biases that could unfairly influence their ratings. Raymond and Reid acknowledge that this is a reasonable concern but state that in their experience they have found that such members provide similar judgments to non-members.

At the standard setting meeting described in this memo, one panelist had extensive experience with the California Bar Exam, having served as the Chair of the Examination Development and Grading (EDG) Team. In my opinion, his experience, coupled with his outspoken nature, may have influenced other panelists. At one point after the facilitator had asked if anyone viewed a certain response being discussed as competent, the panelist exclaimed "You shouldn't even ask if anyone thought the response was competent—don't embarrass them, it can't be competent." I mentioned my concern about the panelist to Elizabeth Parker (Chief Executive Officer) at a break, and she asked Chad if he could gently suggest to the panelist to be aware of the need to let others speak, to which the panelist agreed. My concern was echoed by one panelist's comment on the final evaluation form: "I was troubled that at least one of the panelists had clear familiarity with the existing exam and process and a clear knowledge of 'right' answers as currently graded."

While Raymond and Reid (2001) indicate that in their experience the judgments did not differ across exam committee members vs non-members, in my opinion given the high-profile nature of this standard setting activity, any appearance of bias would best have been avoided by not having any members of the committee involved. However, as Cizek and Bunch (2007) indicated, "the specification of the pool of participants in a standard-setting procedure is first a policy matter. As we have recommended vis-à-vis other policy matters in standard setting, we believe that the entity responsible for setting standards should explicitly discuss and adopt a position on the issue in advance of standard setting" (p. 50). It is my understanding that the policy makers, the California Supreme Court, weighed in in some manner to the effect that having panelists with some direct experience with the Bar Exam would be useful, which is of course their prerogative.

Training of Panelists

The agenda called for an initial training that included purpose and design of the California Bar Exam, performance level descriptor, and Analytic Judgment method. A practice activity with the method followed the training. My comments regarding training on the performance level descriptor is included in the following section, Definition of Performance Standard. Therefore the focus in this section is training provided on the purpose and design of the California Bar Exam and the Analytic Judgment method.

Two slides provided the context for the meeting. The first slide conveyed that the purpose was to "develop a recommended passing score for the California Bar Exam based on the Bar's performance level descriptor" and "communicate recommendation to policymakers who makes the final decision." The second slide indicated that the purpose of the exam, the questions, and grading/scoria criteria or scores could *not* be changed. An additional slide outlined the roles of the attendees. The next slide outlined the steps in standard setting. Three slides followed that focused on the purpose of the bar exam, an overview of the content specifications, and the general scoring framework.

I tend to prefer more text-heavy slides for those panelists who are visual learners, vs. providing limited text and communicating the information verbally. Since there was no evaluation question expressly asking if they understood the purpose or context of the meeting, the only information directly related to this issue was provided by panelists' responses to the open-ended question. The following panelist comments are relevant:

- "More background information before initiating the process would be helpful"
- "It would have been helpful at the top to have a broader discussion about why the study is being done, what the Bar is hoping to learn, and how the individuals (participants) were selected."

There were six slides describing the Analytic Judgment Method. Again, I would have liked to have seen more detail on the slides. For example, "combined results across panelists" is the only text relating to how the cut scores are calculated. Graphics would have been useful for this, as well as other, topics.

Definition of Performance Standard

I noted in my May 8 memo reviewing the workshop plans that the 1-1/2 hours allotted for initial training was, in my opinion, too short. Dr. Buckendahl had indicated that approximately half of that time (45 minutes) would be spent in a conceptual discussion of the definition. The definition would be applied during the practice with the analytic judgment method, which would focus on one essay question in one area. In the memo I conveyed that coming to a common conceptualization of the minimally competent candidate defined by the PLD, and contextualizing it in terms of the exam, is one of the most critical, foundational activities at a standard setting workshop. Ideally, there would be more time planned to be spent in the workshop on applying the definition to the content of the exam. The only application would be during the practice of the method, during which they will also be focused on learning the standard setting task. That discussion would focus on just one area, and so there would be no contextual discussion of the PLD in terms of the other areas. Also, because of the limited amount of discussion that takes place during the workshop, it would be difficult for the facilitator to watch for common errors that standard setting panelists make, such as thinking of the high-achieving candidate or themselves instead of the minimally competent candidate. Ideally there would be more time spent on discussion of the PLD in the context of each of the areas. I noted that, understandably, there are time limitations in terms of the number of days for the workshop, but it is unfortunate that this particular step couldn't be expanded in the agenda.

In the end, more time ended up being taken for the definition of the performance standard than had been indicated on the agenda. This was, in my opinion, a wise decision. Instead of the 45 minutes allotted for this section of the initial training, about 2 hours and 20 minutes was taken. Dr. Buckendahl did an excellent job answering panelists' questions and bringing them back to the concept of the minimally competent candidate. Some panelists said that they would have liked to receive a content outline during the discussion of the minimally competent candidate in the context of content areas. Dr. Buckendahl noted that coming to a common conceptualization of the minimally competent candidate is the hardest part of any standard setting workshop (a statement with which I concur).

Although the discussion of the minimally competent candidate was allotted more time than originally planned for, because of the lack of discussion for most of the questions, Dr. Buckendahl was not able to evaluate panelists' possible misconceptions about the level of knowledge and skills that the minimally competent candidate would display in each content area. It is even more difficult than usual to know if they were using a common conceptualization. If they were not, this would have a direct impact on the ratings they provided.

Also, it is worth noting that a definition of the minimally competent practitioner did not exist until recently. Developing it so close to the time of the workshop, without time for thorough review, may have resulted in a description that is not as well-defined and acceptable to stakeholders as it could be. That was difficult to judge in terms of the workshop activities, however.

Data Collection

According to Kane (1994), there are several strategies that can be used to improve the quality of the data collected in a standard setting study. These include having panelists prove ratings more than once (iteration), engaging the panelists in discussion, sharing performance data with panelists, and providing impact data.

Iteration and Discussion. Kane notes that "if we want any work to be relatively error free, it is generally necessary to review it at least once. Therefore, on this basis alone, iterative procedures, in which the judges get to review their decisions before the passing score is finalized, seems to be preferable to single-pass procedures" (p. 442). The agenda did not call for iteration, and included discussion only during the practice activity. However, Elizabeth Parker asked me on the morning of the second day how much discussion would take place. I told her that there wouldn't be any, and that I had raised that issue in my May 8 memo. After discussion with Dr. Buckendahl, it was agreed that discussion would be incorporated after panelists provided ratings for the first essay. I think this was a good decision, though ideally there would have been discussion, and provision of another round of ratings, for each question. It should be noted that Plake and Hambleton (2001) included iteration and discussion in their inaugural uses of the method. In addition, Zieky, Perie, and Livingston (2008) describe the method as including iteration and discussion accompanied by a full display of the ratings of the panelists for responses.

Performance data. Providing empirical performance data is also viewed as being helpful by Kane (1994) and others (e.g., Jaeger, 1982, 1989). Kane noted that "it is highly desirable that the data collection procedures promote consistency in the data being generated. One way of doing this is to have the judges discuss their ratings after they have independently judged the items. Fitzpatrick (1989) has pointed out some potential problems associated with group dynamics that we need to be concerned about, but the benefits of having the judges consider their judgments as a group seem to outweigh the risks" (p. 442). Empirical data were not provided during this workshop. Although conceivably panelists could have been given information about the distribution of scores obtained for a given question, that would not have been advisable for the Analytic Judgment method given that the panelists might then stray from thinking about the minimally competent candidate and make a more normative judgment.

Consequence data. Another type of data that is sometimes provided are consequence data. As Kane notes, "If they are available, external checks on the judgments being made would also be helpful. There is no good reason to ignore information about the consequences of the decision, if such information is available (Busch & Jaeger, 1990; Linn, 1978; Norcini, Shea, & Kanya, 1988). In many situations, we are forced to make decisions without knowing much about the consequences, but, if good data on the probable consequences of decisions were available, it would seem to be prudent to use it and irresponsible to ignore it. In particular, data on the consequences of setting the passing score at different points may be useful in helping judges to make realistic decisions" (pp. 442–443). Consequences data were not provided during this workshop. Reasons may have included the fact that the current pass rates are getting a lot of attention, and making the panelists aware of the pass rates may have been viewed as being too risky lest they change their ratings to change the cut score and thus the pass rate to be more like those of other states. In addition, estimating the pass rates involves equipercentile equating, which may have proved challenging to do at the workshop as the cut scores were being calculated.

Scoring rubrics. A factor not specifically described by Kane (1994), but in my opinion related to the quality of the data collected, is that of scoring rubrics. A topic of great debate among the deans at their April 6 meeting, and likely in other venues, was whether to provide scoring rubrics to the panelists. They voiced concern that providing them could bias the panelists' judgments and downplay the expertise they are bringing as a result of their experience in the field.

However, familiarizing panelists with rubrics is a standard practice in standard setting. The purpose is to inform panelists of the dimensions along which the responses will be evaluated. Panelists are not shown the scores, just the rubrics. As I noted in my May 8 memo, Plake and Hambleton (2001), in their description of the first implementations of the Analytic Judgment method, describe that panelists were given "training on the questions and scoring rubrics" (p. 291). I was able to speak with both authors in April, 2017 at a professional meeting, and asked them for their general opinion on whether familiarization with the rubrics would be advisable in an implementation of the Analytic Judgment method, and they both strongly affirmed that it would. Their concern, which I share, is that panelists would have no context for the kind of information that a strong response would contain, and as a result may introduce their own idiosyncratic ideas in the process.

In a discussion with Dr. Buckendahl before the workshop, he indicated that the scoring rubrics range from 1 to 6 pages in length. Model essay responses are released, though with no rubrics. He indicated that he was planning to describe a core set of skills that are evaluated for each of the essay questions and performance task. These core skills would be: issue spotting, identification of elements of the applicable law/legal principle, analysis of the law as applied to the facts, conclusions on likely outcomes, and justification/reasoning for the conclusion. I concurred that at the least, these core skills be explained to the panelists so that they would be aware of the dimensions on which the responses are being evaluated. These skills were in fact communicated to panelists as the general scoring framework.

The report notes that specific rubrics were not provided in order to "avoid potentially biasing the panelists in their judgments and to focus on the common structure of how the constructed response questions were scored" (Buckendahl, 2017, p. 1). I do not agree with this statement, and based on what I observed at the workshop, I continue to believe that providing rubrics would have been beneficial. Although panelists are often concerned that they are not qualified to provide judgments, I think that not being able to get at least of a sense of what would make a good response caused them a fair degree of anxiety and to not have great confidence in their ratings. Panelists' comments included the following relevant to scoring rubrics; three comments indicated that they would have been helpful, while two comments indicated that not providing rubrics was appropriate.

- "I'm still not completely certain that I understand how we are qualified to do this without answers. It seems like this could have the overall effect of making it easier to pass?"
- "I am concerned that an unprepared attorney, without the benefit of experience, studying, or a rubric, is not a good indicator of a minimally competent attorney"
- "It might be helpful to have some kind of 'correct' sample answer to avoid having to go back and re-score or re-read for lack of knowing "the correct answer"
- "Although providing a scoring rubric would make categorization more consistent, it would do so in view of the thoughts of the author and not of the 20 panelists. Having no rubric was tough, but appropriate"
- "The performance test, unlike subject matter knowledge tests (essays) is much more amenable to this sort of standard setting. While, as with essays, we did not outline/rubric/calibrate, that is less necessary because of closed universe and the skills being tested"

5) Panelist Feedback

Design of Evaluation Forms

The original agenda had panelists completing two evaluation forms, with the first being given at mid-day on the first day of the workshop, after the practice activity, and the second at the end of the workshop. In my May 8 memo I suggested that the first evaluation form have panelists indicate on the first survey, done after the practice activity, whether they felt confident in their ability to perform the task so that remediation could be done before rating/judging of the "real" essays. I also strongly suggested having additional evaluations, such as at the end of the first day when they will have completed ratings of the first essay, in order to allow for targeted remediation the next morning. Showing panelists that evaluation feedback is being reviewed and addressed can increase their confidence in the process as well as answering any questions they may have. I also suggested having an evaluation at the end of the second day.

There were four evaluation forms administered during the workshop, which was a useful modification to the design. However, the forms were rather short, and I would have liked to have seen additional questions. For example, on the first form a question asking how clear they were with the purpose of the meeting would have been informative. Also, questions directly asking about their confidence in their understanding of (1) the minimally competent candidate and (2) the rating task on each evaluation form would have allowed for a comparison of how their confidence grew, if at all, over the course of the workshop. Because of the limited amount of discussion, a question asking whether they thought the discussion was valuable would have shed light on whether it would have been good to include more opportunities for it.

Since panelists were not informed of the recommended cut score during the workshop, no question could be asked about their confidence in that outcome. On the final evaluation form panelists were asked "Overall, how would you rate the success of the standard setting workshop?" However, it is not clear to me what the panelists would view as "success." I would have liked to see a question asking something along the lines of "How confident are you that the activities conducted during the workshop would yield a recommendation of an appropriate and defensible cut score?" This question reflects the fact that they do not know the recommended cut score, but asks about whether they think the workshop activities would yield one.

Question 1 on the first evaluation form also asked about success of four activities: orientation to the workshop, overview of the exam, discussion of the PLD, and training on the methodology. Again, I am not sure how the panelists would define success and if they would define it similarly.

The ratings scales used for the same category of questions also had some inconsistencies, making it difficult to make valid comparisons across questions in the same category:

- "Success" questions: Question 1 on form 1 had labels only for the ends of the rating scale (i.e., very unsuccessful and very successful), whereas question 14 on the final evaluation form had all four points on the scale labeled (very unsuccessful, unsuccessful, successful, and very successful).
- "Time" questions: Question 2 on form 1 had three points on the scale (too little time, the right amount of time, and too much time). The other questions related to time (questions 4, 7, 10, and 13) had four points on the rating scale (not enough time, barely enough time, sufficient time, more than enough time).
- "Confidence" questions: The questions asking about confidence had the label "somewhat confident" for rating point 3. Other questions (with the exception of questions 2 and 3) had an unqualified label for that rating point, such as "sufficient," "successful," and "organized." It would have been more consistent to have rating point 3 be labeled as "confident" rather than "somewhat confident."

Because of these rating scale inconsistencies, comparisons of median ratings across the questions could be misleading. More appropriate interpretations could be made by looking at the frequency of responses to each questions and within the same category (with the exclusion of questions 1 and 2, which had different labels than the other questions in that category).

Feedback Obtained

Most of the questions on the evaluation forms fell into three categories: timing, confidence, and success. It is of interest whether any panelists provided negative responses to any of the questions (the report provides the full range of responses to each question; the breakdown of frequency between the two positive responses is not provided here).

Timing questions were asked about training, practice ratings on the first day, and operational ratings on each of the three days. For every question except day 1 operational recommendations, some panelists reported that they had not had enough time. Four panelists thought there was too little time allotted to training. For practice and operational ratings on days 2 and 3, one or two panelists thought there was not enough time, and between one and six panelists thought there was barely enough time.

Four questions asked about panelists'' confidence levels; one question asked about moving from practice to operational ratings, and three about operational recommendations made each day. In general, one expects panelist confidence to grow over the course of the workshop, which was the case for the most part. One panelists felt not at all confident moving from practice to operational, and one felt not very confident. Confidence in day 1 recommendations had one panelist not at all confident, and 2 not very confident. For day 2, no panelists indicted not at all confident, and one indicated not very confident. For day 3, no panelists indicated not at all confident or not very confident. This is a positive outcome.

Several questions asked about the success of different activities conducted during the workshop. One question asked about four different aspects of the training—orientation, overview of the exam, discussion of the PLD, and training on the methodology. No panelists rated any of the activities as very unsuccessful. One and two panelists gave a rating of 2 (which had no label) to discussion of the PLD and training on the methodology, respectfully. One question asked about the overall success of the workshop; no panelist rated it as very unsuccessful, and one panelist rated it as unsuccessful.

One question that does not fall into the previous categories asked about the overall organization of the workshop. No panelists rated it as either very unorganized or unorganized.

Panelists were also asked open-ended questions on each of the four evaluation forms. (Because the forms were anonymous, it is not possible to know whether the same or different panelists provided responses on different forms throughout the process.)

- Several panelists praised the facilitation skills of Dr. Buckendahl:
    - "Dr. Buckendahl trained us very effectively. He is engaging, clear, and attentive. I have confidence in him and the process. Good work!" (evaluation form 1)
    - "Work with Dr. Buckendahl again. He was very careful, clear, and engaging. Well done!" (evaluation form 4)

- There were comments on each of the forms about panelists feeling rushed and not having enough time, in line with the responses provided to the rating-scale questions (particularly the six panelists who said there was barely enough time for practice):
    - "Had to rush in order to have time for lunch" (evaluation form 1)
    - "I did not finish and felt rushed. More time for first question" (evaluation form 2)
    - "It was very difficult to read 60 essays in one day" (evaluation form 3)
    - "I really found the time available to review the subject-matter answers to be very challenging" (evaluation form 4)
    - "I had a hard time with the time limit to review each answer. I am not clear if I was being too thorough, or I missed the lesson on how to move through answers at a quicker pace." (evaluation form 4)
- There were several comments voicing concern about the panelists' not feeling qualified to do the task:
    - "Not convinced this methodology is valid. Many of us clearly do not know some applicable law and these conclusions may therefore determine that incompetent answers amounting to malpractice are nevertheless passing/competent." (evaluation form 1)
    - "I am concerned that an unprepared attorney, without the benefit of experience, studying, or a rubric, is not a good indicator of a minimally competent attorney. We all have an ethical duty to become competent. New lawyers/3 Ls do that by preparing for the exam. A more seasoned lawyer does that by refreshing recall of old material or by resort[ing] to practice guides. Having neither the benefit of studying nor outside sources, at least some of us may be grading with lack of minimum adequate knowledge. By studying for the exam, test-takers are becoming competent and gaining that minimal competency. Practicing professionals who become specialized may lose/atrophy that competence in certain field, which needs to be refreshed by CLG and other sources. So these scores may be of limited utility." (evaluation form 2)

6) Documentation

This criterion refers to extent to which features of the study are reviewed and documented for evaluation and communication purposes. As I noted above under the explicitness criterion, I think more detail could have been provided in the plans. However, the report does include more detail, which means that the method can be clearly understood and, if necessary, replicated based on the documentation. The report is structured around the same framework I have used in this memo—procedural, internal, and externality sources of validity evidence, which is useful.

Internal Evidence

Evidence discussed within this section includes consistency within method, and intrapanelist and interpanelist consistency.

1) Consistency Within Method

Kane (1994, 2001) noted that the best way to estimate the standard error of the passing standard is to convene different groups of panelists on the same or different occasions to implement the same method. As if often the case, resources did not allow for such an approach in the case of the California Bar Exam. An alternative approach is to calculate the standard deviation of the panelists' cut scores. The report does provide an estimate of variability in the standard errors of the mean and median. I would have liked to see the report contain a table with panelist-level cut scores so that these calculations could be replicated, however.

2) Intrapanelist Consistency

Evidence for this criterion is provided by the degree of relationship between each individual panelist's ratings and empirical data provided to them. Since no empirical data were provided, this criterion is not relevant.

3) Interpanelist Consistency

Evidence for this criterion is provided by the degree to which question-level ratings were consistent across panelists, both within and across rounds. The report did not contain question-level cutscores (or, as noted previously, panelist-level cut scores), so these data were not available for review.

External Evidence

1) Comparisons Between Methods

Kane (2001) noted that if two standard-setting methods are used, "if the two approaches agree, we have more confidence in the resulting cutscores than we would have if either method were used alone" (p. 75). The meeting did not afford the opportunity to directly compare the results of two standard-setting methods. This is not unusual, since comparing the results of two methods is not often done operationally; it is more often undertaken when a new method is being compared to an existing one.

2) Reasonableness of Passing Standard

Evidence relating to the reasonableness of the cut score obtained is provided in this case by impact data, specifically the percentage of candidates estimated to pass the exam using the recommended cut score. The report does a very good job of describing the results in terms of the mean and median cut scores, and notes that since the mean and median did not converge, the median is a more appropriate measure of central tendency since it is not as sensitive to outliers. The report also presents cut scores that would result if we took into account error—specifically 1 or 2 standard error of the median above and below the mean and median cut scores.

The median, panel-recommended cut score is effectively the same as the current cut score, and therefore little change in pass rates would result. This pass rate is much lower than most other states (National Conference of Bar Examiners, 2017), which has resulted in debate among stakeholders. However, as the report notes, comparisons of California's passing rate to those of other jurisdictions should be done with caution since each state has different eligibility criteria and definitions of minimum competency. The report also describes California's more inclusive policies and the likely impact on the candidate pool and pass rate, as well as downwards trends in pass rates across the country.

A key concept described in the report is that of classification error—false positives and false negatives—and how the standard errors can be used to lower or raise the cut score to minimize one at the expense of the other (Cizek & Bunch, 2007). This is an important consideration for policy makers to take into account as they review the panel recommendation, and the report does a good job of relating potential adjustments to the cut score to the two types of errors.

Conclusion

As Zieky et al. (2008) note, cut scores cannot be categorized as right or wrong, since the "rightness" of a cut score is a matter of values, and different stakeholders have different values. For that reason, we must rely upon the accumulated weight of the various sources of validity evidence outlined in the previous sections of the memo. Because although validity evidence cannot establish that a cut score is appropriate, the lack of such evidence can cast doubts on its validity.

The validity evidence summarized in this memo reveals some shortcomings. In my opinion the method could have been better implemented, for example providing score rubrics, including more than one round of ratings, and including discussion. Some of the implementation features may have been due to time constraints, i.e., the length of the workshop. Given those limitations, Dr. Buckendahl did a very good job of implementing the method, including conducting the training and facilitating other activities.

Although in some ways the design and execution fell short of what I would view as best practice, in my opinion there were no fatal flaws. The panel-recommended passing score, and the possible approaches to adjusting it made by Dr. Buckendahl, represent credible information for the Supreme Court to consider when they make their policy decision. An observation by Kane (2001) provides a fitting conclusion:

> The policy interpretation cannot be directly evaluated in terms of empirical data. Empirical data cannot tell us how much of a good thing is enough. The appropriate policymaking body must set the standards by deciding on an appropriate balance among competing goals. (pp. 80—81)

References

Berk, R. A. (1986). A consumer's guide to setting performance standards on criterion-referenced tests. *Review of Educational Research, 56,* 137–172.

Buckendahl, C. W. (2017, July 28). *Conducting a standard setting study for the California Bar Exam: Report*. Las Vegas, NV: ACS Ventures.

Busch, J. C., & Jaeger, R. M. (1990). Influence of type of judge, normative information, and discussion on standards recommended for the National Teacher examinations. *Journal of Educational Measurement, 27,* 145.

Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests.* Thousand Oaks, CA: Sage.

Fitzpatrick, A. R. (1989). Social influences in standard-setting: The effects of social interaction on group judgments. *Review of Educational Research, 59,* 315–328.

Hambleton, R. K. (2001). Setting performance standards on educational assessments and criteria for evaluating the process. In G. Cizek (Ed.), *Standard setting: Concepts, methods, and perspectives* (pp. 89–116). Mahwah, NJ: Erlbaum.

Hambleton, R. K., & Pitoniak, M. J. (2006). Setting performance standards. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 433-470). Westport, CT: Greenwood/Praeger.

Hambleton, R. K., & Plake, B. S. (1995). Using an extended Angoff procedure to set standards on complex performance assessments. *Applied Measurement in Education, 8,* 41–55.

Jaeger, R. M. (1982). An iterative structured judgment process for establishing standards on competency tests: Theory and application. *Educational Evaluation and Policy Analysis, 4,* 461–475.

Jaeger, R. M. (1989). Certification of student competence. In R. Linn (Ed.), *Educational measurement* (3rd ed., pp. 485–514). Englewood Cliffs, NJ: Prentice-Hall.

Kane, M. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research, 64,* 425–461.

Kane, M. (2001). So much remains the same: Conception and status of validation in setting standards. In G. Cizek (Ed.), *Standard setting: Concepts, methods, and perspectives* (pp. 53–88). Mahwah, NJ: Erlbaum.

Kingston, N. M., Kahl, S. R., Sweeney, K., & Bay, L. (2001). Setting performance standards using the body of work method. In G. J. Cizek (Ed.), *Standard setting: Concepts, methods, and perspectives* (pp. 219–248). Mahwah, NJ: Erlbaum.

Linn, R. L. (1978). Demands, cautions, and suggestions for setting standards. *Journal of Educational Measurement, 15,* 301-308.

National Conference of Bar Examiners. (2017, March). 2016 Bar Examination and admission statistics. *The Bar Examiner, 86*(1).. Retrieved from http://www.ncbex.org/pdfviewer/?file=%2Fdmsdocument%2F205

Norcini, J., Shea, J., & Kanya, D. (1988). The effect of various factors on standard setting. *Journal of Educational Measurement, 25,* 57-65.

Olson, B., Mead, R., & Payne, D. (2002). *A report of a standard setting method for alternate assessments for students with significant disabilities* (Synthesis Report 47). Minneapolis: University of Minnesota, National Center on Educational Outcomes. Retrieved from http://education.umn.edu/ NCEO/OnlinePubs/Synthesis47.html

Plake, B. S., & Hambleton, R. K., (19XX). The analytic judgment method for setting standards on complex performance assessments. In G. J. Cizek. (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 283-312). Mahwah, NJ: Erlbaum.

Raymond, M. R. & Reid, J. B. (2001). Who made thee a judge? Selecting and training participants for standard setting. In G. J. Cizek (Ed.), *Standard setting: Concepts, methods, and perspectives* (pp. 119–157). Mahwah, NJ: Erlbaum.

State Bar of California. (2016, November 18). *July 2016 Bar Exam results: California Bar Exam pass rate summaries.* Retrieved from http://www.calbar.ca.gov/About-Us/News-Events/News-Releases/ArtMID/10234/ArticleID/140/State-Bar-announces-results-for-July-2016-California-Bar-Examination

State Bar of California. (2017, May 12). *February 2017 Bar Exam results: California Bar Exam pass rate summaries.* Retrieved from http://www.calbar.ca.gov/About-Us/News-Events/News-Releases/ArtMID/10234/ArticleID/140/State-Bar-announces-results-for-July-2016-California-Bar-Examination

van der Linden, W. J. (1995). A conceptual analysis of standard setting in large-scale assessments. In *Proceedings of the joint conference on standard setting for large scale assessments of the National Assessment Governing Board (NAGB) and the National Center for Education Statistics (NCES), Volume II* (pp. 97–117). Washington, DC: U.S. Government Printing Office.

Zieky, M. J., Perie, M., & Livingston, S. (2008). *Cutscores: A manual for setting standards of performance on educational and occupational tests.* Princeton, NJ: Educational Testing Service.

Appendix

Hambleton's 20 Criteria for Evaluating a Performance Standard-Setting Study

The following 20 criteria were presented by Hambleton (2001). Many of them overlap with issues considered in the validity framework presented in the body of the memo, based on Kane's (1994, 2001) formulation. At the request of the Bar, brief notes are presented here about how the standard-setting process for the California Bar Exam would be evaluated given these criteria.

Hambleton Criterion	Criterion/Section in Memo	Notes
1) Was consideration given to the groups who should be represented on the standard-setting panel and the proportion of the panel that each group should represent?	• Procedural ○ Implementation of procedures – Selection of panelists	• Length of practice and setting (with separate category for faculty/educator) were the only demographics considered • Clear targets for the number of panelists in each category were not documented
2) Was the panel large enough and representative enough of the appropriate constituencies to be judged as suitable for setting performance standards on the educational assessment?	• Procedural ○ Implementation of procedures – Selection of panelists	• The number of panelists was sufficient • Representativeness cannot clearly be evaluated due to the lack of specific documented targets
3) Were two panels used to check the generalizability of the performance standards across panels? Were subpanels within a panel formed to check the consistency of performance standards over independent groups?	• Internal ○ Consistency within method	• Only one panel was used • The panel was not split into subpanels • The standard error of the mean and median were reported; however, panelist-level cut scores were not included to allow for replication of the calculations
4) Were sufficient resources allocated to carry out the study properly?	—	• Most resources allocated were sufficient • An adequate number of panelists was recruited • However, ideally more time would have been allotted for the workshop to allow for more time for ratings and discussion
5) Was the performance standard-setting method field tested in preparation for its use in the standard-setting study, and revised accordingly?	• Procedural ○ Selection of procedures	• No field testing was done

Hambleton Criterion	Criterion/Section in Memo	Notes
6) Was the standard-setting method appropriate for the particular educational assessment and was it described in detail?	• Procedural ○ Selection of procedures ○	• The method was appropriate for the exam • Other possible methods were mentioned but not described in the report • The form of the method used in this workshop should have been described in more detail in the plan, but that was rectified in the report
7) Were panelists explained the purposes of the educational assessment and the uses of the test scores at the beginning of the standard-setting meeting? Were panelists exposed to the assessment itself and how it was scored?	• Procedural ○ Implementation of procedures – Training of panelists	• Panelists were given an overview of the purpose of the exam • Panelists were given a list of the high-level scoring criteria • Detailed rubrics were not shared with the panelists
8) Were the qualifications and other relevant demographic data about the panelists collected?	• Procedural ○ Implementation of procedures – Selection of panelists	• I did not see documentation of the qualifications and demographic data other than years of experience and setting; however, it may well exist
9) Were panelists administered the educational assessment, or at least a portion of it?	• Procedural ○ Implementation of procedures – Training of panelists	• Panelists did not take the exam under operational conditions • The agenda called for them to outline responses to the essays and performance task, but this was not done in practice
10) Were panelists suitably trained on the method to set performance standards? For example, did the panelists complete a practice exercise?	• Procedural ○ Implementation of procedures – Training of panelists	• The panelists did receive training on the method • Panelists did complete a practice exercise
11) Were descriptions of the performance categories clear to the extent that they were used effectively by panelists in the standard-setting process?	• Procedural ○ Implementation of procedures – Training of panelists	• The definition of the minimally competent candidate was reviewed with the panelists • More time could have spent on discussing the minimally competent candidate in the context of the each of the essay questions and the performance task

Hambleton Criterion	Criterion/Section in Memo	Notes
12) If an iterative process was used for discussing and reconciling rating differences, was the feedback to panelists clear, understandable, and useful? Were the facilitators able to bring out appropriate discussion among the panelists without biasing the process?	• Procedural ○ Implementation of procedures – Data collection implemented in systematic and thorough fashion	• There was limited discussion, only in the practice round and for the first essay question • The facilitator did an excellent job of facilitating discussion
13) Was the process itself conducted efficiently? Were the rating forms easy to use? Were documents such as examinee booklets, tasks, items, and so on, simply coded? If copies of examinee work were being used, were they easily readable? Were the facilitators qualified?	• Procedural ○ Implementation of procedures – Data collection implemented in systematic and thorough fashion	• All materials were color-coded and easy to follow • The facilitator was highly qualified and effective
14) Were panelists given the opportunity to "ground" their ratings with performance data, and how was the data used?	• Procedural ○ Implementation of procedures – Data collection implemented in systematic and thorough fashion	• No performance data were provided to panelists
15) Were panelists provided consequential data (or impact data) to use in their deliberations, and how did they use the information? Were the panelists instructed on how to use the information?	• Procedural ○ Implementation of procedures – Data collection implemented in systematic and thorough fashion	• No consequence data were provided to panelists

Hambleton Criterion	Criterion/Section in Memo	Notes
16) Was the approach for arriving at final performance standards clearly described and appropriate?	• Procedural ○ Implementation of procedures – Training of panelists	• The approach was appropriate • It was not well-described to panelists • It was summarized in sufficient detail in the report
17) Was an evaluation of the process carried out by the panelists?	• Procedural ○ Panelist feedback	• The panelists completed four evaluation forms • There were some issues with consistency of the rating scales • Additional questions would have been useful
18) Was evidence compiled to support the validity of the performance standards?	• Procedural ○ Documentation	• The report summarizes evidence in relation to the three sources of validity evidence
19) Was the full standard-setting process documented (from the early discussions of the composition of the panel to the compilation of validity evidence to support the performance standards)?	• Procedural ○ Documentation	• Early documentation in terms of plans could have included more detail • The report contained sufficient detail
20) Were effective steps taken to communicate the performance standards?	• Procedural ○ Documentation	• Documentation contained sufficient detail • Communication of the cut score to policy makers or other constituencies was not provided to me for review